Discussion

Daniel Kasprzyk
National Center for Education Statistics

The meetings of the American Statistical Association provide an excellent forum for data collection programs to discuss their findings and methods. Many survey programs benefit from the discipline involved in preparing research papers and presentations for these meetings. The National Science Foundation(NSF) has taken the opportunity on several occasions over the last few years to bring the statistical community up-to-date on the progress of their data system on scientists and engineers.

As discussant in 1992 of several NSF papers describing research aimed at improving their data collection system(Kasprzyk, 1992), I was impressed with the program's scope of activities. Staff was ambitious and energized by the tasks. Much of the work was directed to identifying and improving the content of the survey system as well as its response rates. Their success is apparent in the improved rates of response in all components of the data collection system. The manuscripts in this session are the natural next stage in the evolution of developing a body of information about the quality of the data in the NSF data system.

This stage is important, but can easily be dismissed because of higher priorities(develop the next cycle of questionnaires, release data etc.) or lack of funding. So it is a pleasure to see findings from the NSF surveys presented again at these meetings.

This session presents results from each of the three demographic NSF surveys:

1. Survey of Doctorate Recipients(SDR), conducted by the National Research Council. Using a list of all U.S. doctorate holders, the data collection had both a mail survey component as well as a computer-assisted telephone interviewing follow-up.
2. National Survey of Recent College Graduates(NSRCG), conducted by Westat Inc. Using a two stage sample, first of institutions offering bachelor's and master's degrees in science and engineering and then a sample of degree recipients in science and engineering for graduation years 1990, 1991, 1992, Westat interviewed the sampled individuals using Computer Assisted Telephone Interviewing(CATI).
3. National Survey of College Graduates(NSCG), conducted by the Census Bureau. Using a list developed from the decennial census, the Census Bureau drew a sample of persons receiving at least a bachelor's degree before 1990. The mode of administration for this survey was first a mail-out/mail-back questionnaire, computer-assisted telephone interviewing(CATI) for mail nonrespondents, and finally a personal visit, if necessary.

This is a complicated data system where each component has its own design parameters. Thus, each system literally requires its own research and evaluation program. This is a formidable challenge for those responsible for the program. So it behooves them to identify important studies that have cross- component implications.

The Hardy, Mooney, and Eisenhower manuscript provides the much needed overview of the NSF research program. Mitchell et al tackle the issue of nonrespondents in the Survey of Doctorate Recipients and Tremblay and Moore study nonrespondents in the NSCG. The McGuinness et al manuscript look at interviewer effects in the NSRCG and NSCG surveys with a view to, perhaps, improving the questionnaire and interviewer training.

Some general remarks on each of the four papers follow below.

Hardy, Mooney, and Eisenhower

The need for an ongoing research and evaluation program has not diminished over the years. The authors provide general principles to drive the research and evaluation program:

1. a broad array of research should be conducted, both statistical and cognitive;
2. research projects are defined relative to their contributions in resolving questionnaire or procedural problems or to the information provided to data users;
3. the skills and interests of the participating organizations are used;
4. complementary projects over-time and across organizations should be implemented.

These are useful criteria, but let me add a few more. First, I think it is important to maximize the use of research findings across the various components. This

may be implicit in the Hardy et al criteria, but it needs to be restated because of the cost and staffing implications. There are enough survey design differences that any cross component findings may be only suggestive of results across components. So some caution needs to be exercized.

Second, another approach to designing a program of research and evaluation is to develop a quality profile for the data collection program. This should lead to the identification of issues requiring further work. In other words, the quality profile defines the research agenda not the other way around. I have been associated with two quality profiles(Jabine, King, and Petroni, 1990; Jabine, 1994). The Survey of Income and Program Participation(SIPP) model(Jabine et al, 1990) documents research results but does not address next steps in the research program. In the Schools and Staffing Survey(SASS) model(Jabine, 1994), explicit recommendations are found in an internal memorandum, which we expect to be made available as a supplement to the quality profile. The important point is that the quality profile identifies areas in the survey program requiring further work.

Third, to what extent has the research agenda been developed based on the recommendations of the National Academy of Sciences Study Panel(Citro and Kalton, 1989)? I have not reviewed the panel's recommendations recently, but it is very important that the issues the panel raised are addressed.

Fourth, identifying research areas likely to improve methods, procedures, or data in the next collection cycle should be given high priority.

Finally, it is nice to see a significant amount of time and energy being given to a data collection program on a continuous basis rather than simply at the time of a major redesign.

Mitchell, Moonesinghe, and Pasquini

Nonresponse in surveys is a popular topic at these meetings. Even simple descriptive analyses of nonresponse can present interesting, although incomplete information. Why is nonresponse so interesting and popular a topic? First, the problem of nonresponse or nonresponse bias is one of the few aspects of nonsampling error where it is easy to develop suggestive indicators of potential bias calculating using response rates and estimates of differences in the characteristics of respondents and nonrespondents. Second, attempts at characterizing differences between respondents and nonrespondents help the data user in his/her assessment of the quality of the survey data relative to his/her application. Third, occasionally nonresponse analyses identify a problem in the data collection procedures and be suggestive of where the procedures can be improved. Fourth, understanding which variables distinguish respondents from nonrespondents can help in determining the appropriate nonresponse adjustment cells in the survey system's estimation procedures. Fifth, survey nonresponse is also a topic where most people have a point of view as to how to reduce its magnitude, thereby reducing the potential for nonresponse bias.

Data collection programs, at a minimum, report unit nonresponse rates. Additional analyses of nonresponse are completed infrequently. So it is with pleasure that we can point to the two papers in this session as furthering the knowledge base on nonresponse in the NSF data system.

The Mitchell et al paper compares characteristics of respondents - early mail returns, interim mail returns, and late telephone interviewing follow-up - through the use of variables on the sampling frame(field of doctorate, sex, race, and age). The goal of this analysis is to determine whether a release of early estimates from the survey is possible.

Several questions come to mind with these analyses:

1. The need for some multidimensional tables is apparent. These tables are possible because of the large sample and would help differentiate respondents and nonrespondents.
2. The intent of the analysis is to develop early releases of the data by using the early mail returns as an estimate for the full sample estimate. This is an important question. To answer the question, more needs to be known about the importance of estimating levels, such as the number of employed, rather than simply estimating proportions. Furthermore, it is impossible to say much about the differences in characteristics without having some idea of the magnitude of the sampling error, and these are not shown.
3. Efforts to improve timeliness are laudable, but the need for the early release does depend on the application and user requirements. Is the issue a set of "early release" tables or will it be an "early release" microdata set? If the release is the latter, be aware that a substantial amount of work may go into the early release; it is, after all, a new data set.

Clearly, the study of early respondent estimates and characteristics of nonrespondents to the Survey of Doctorate Recipients(SDR) has just begun. The research

community anticipates seeing additional detail in the future.

Tremblay and Moore

The Tremblay and Moore paper focusses on nonresponse in the National Survey of College Graduates(NSCG). The authors use sampling frame data collected on the decennial census long form; however, some attention is also paid to the data collection mode ultimately used to obtain the interview. Several of the issues I mentioned previously are the goals of this research. First, it is desirable to provide some information about survey nonrespondents, since little appears to be known about the NSCG nonrespondents. Furthermore, it is not obvious what variables are important in predicting nonresponse; these variables may be helpful in developing a nonresponse adjustment strategy. Second, the issue of nonresponse over time is important, since each survey's sample serves as the sampling frame for the next data collection cycle. Nonresponse bias at each data collection cycle could affect estimates at a future collection cycle. An important issue the survey program wants to address is whether to followup on each cycle's nonrespondents in future collection cycles. Finally, the data collection program needs to determine the operational and fiscal merits of using a uniform collection mode across all subgroups; that is, the authors suggest it may be more practical and less expensive to determine in advance that certain subgroups are candidates for a particular mode.

The end result of the paper is disappointing because the thirteen frame variables appear to be of little help in characterizing respondent/nonrespondent differences. An almost similar result occurs in NCES Schools and Staffing Survey(Salvucci, Monaco, Gruber, 1995), however, in that analysis a few variables do emerge. Another nonresponse bias analysis I have seen recently concerned the NCES RDD National Household Education Survey(Brick, 1995). Here again variables available for the nonresponse bias analysis indicated no apparent bias. Of course, these nonresponse analyses can only provide indications of nonresponse bias. They can not tell a typical analyst whether his/her results are biased, because analysts always use variables for which no information is available on the nonrespondents. Thus, taking a cautious approach with respect to decisions concerning nonresponse is taking the wise approach.

With this in mind, I offer these general comments and questions:

1. I am uncomfortable with the important conclusion that characteristics of respondents and nonrespondents do not differ. Why? Because, first, I do not know the CART methodology to render a judgement about its proper implementation. The selection of the CART samples appears to be simple random, but shouldn't they mirror the actual sample design, which, I believe, has differential selection probabilities. In the same spirit, the unweighted CART analysis requires stronger justification. Second, I prefer to compare both weighted and unweighted response rates and distributions of respondents and nonrespondents. The weighted analysis, because of the different selection probabilities, allows us to test univariate and multivariate distributions of respondents and nonrespondents. Third, having developed the weighted estimates, simple comparisons of survey aggregates with other sources, such as the Current Population Survey, seem advisable to establish the face validity of the data.
2. Establishing that respondents and nonrespondents do not differ on a number of frame variables is an important finding for deciding to keep only respondents on the sampling frame for the next data collection cycle, but I need stronger justification to accept the finding. Other longitudinal surveys, although they have different goals, return to nonrespondents on the wave following their nonresponse. The National Longitudinal Survey of Youth(NLSY) comes to mind. During the last several years, the Panel Study of Income Dynamics(PSID), after not returning to noninterview households for many years, has changed its policy, subject to the availability of funding. The Survey of Income and Program Participation(SIPP) used a rule - two nonresponses in succession - to decide when to stop trying to interview a nonrespondent. Furthermore, research exists to suggest that respondents and nonrespondents can differ on important variables. The SIPP has spent an enormous amount of time, money, and energy to develop better nonresponse adjustment procedures to help compensate for these known differences. The **Proceedings** for the last several years contain several manuscripts on these topics.So I take a fairly cautious point of view concerning the exclusion of nonrespondents in the current sampling frame. I would want a stronger case made; I might even suggest maintaining a nonresponse stratum and drawing a small sample during each future collection cycle for the express purpose of again doing a nonresponse bias study. Why should anyone expect the finding, if it holds up under further scrutiny, to remain the same over time? Shouldn't the sponsors

of this survey program stay on top of the issue over time?

3. I do not understand the need to include the out-of-scope cases in this study. It seems to me that out-of-scope cases are out of the analytic universe for the study. If properly identified through the screening procedure, they have no role in the analysis. However, if the survey program thinks too many cases are being identified as out-of-scope, then, perhaps, the procedure to identify scientists and engineers ought to be studied and revised.
4. The authors suggest that further research may help the program operations staff determine a more efficient assignment of sample cases to a data collection mode, suggesting that some subgroups may benefit by being assigned to CATI rather than to mail. I am told when this kind of assignment is done it can save money. It is important for all concerned with this issue to obtain better cost data (not a Census Bureau strength) in order to make a more informed judgement. This is a situation, as it is in all surveys, where cost is not the only issue; the quality of the data collected is also very important. This facet of the problem is not addressed in this manuscript and clearly more research is needed to balance the trade-offs. Again, this is a situation where more evidence on the subject is desirable. It seems to me that if money is the issue, and it always is, I would try to maximize the mail return rate and minimize the number of personal visits. Clearly, everyone needs to know more.

McGuinness, Brick, Lapham, Cahalan, and Owens

The McGuinness et al paper is concerned with measuring interviewer effects to help identify items requiring improved question wording. It uses two models, one for each of the surveys, NSRCG and NSCG, to estimate interviewer effects. The paper also summarizes a study of a small number of interviewers and a small number of items to determine 1) whether large interviewer effects are correlated with items requiring a substantial amount of probing and 2) whether items requiring more than the average amount of probing have large interviewer effects.

The authors acknowledge the cost problems of implementing a design that has an interpenetrating sample, a design well-suited to answer questions about interviewer effects; they also are quite honest in discussing the limitations of the analyses. The paper is indicative of what we will see in the future. As budgets decrease, costly experimental designs and reinterview programs to clarify and understand response issues will occur less frequently. Analyses, like the ones described in this paper, will become more prominent, even though a significant number of model assumptions may be violated.

The authors have provided a good discussion of the interviewer effects models, their results, and the differences in the surveys used to develop the models. Given the large number of differences between the two data collection programs, it is remarkable that the findings across the two surveys are similar. That is, large interviewer effects are observed in the same or similar items across the two surveys. The magnitude of the effects appears to be different, but this, it seems, is less important than the finding that similar items may be problematic in the surveys.

The conclusion then is fairly obvious - that the identified items ought to be revised and retested prior to the next data collection cycle. The authors say, however, that the most common feature of the problem items was that they were primarily open-ended questions. This is not a new finding(Bushery, Royce, and Kasprzyk, 1993). We, as a community, may be relearning the obvious. If we want to reduce large interviewer effects and large response variance, then we ought to reduce the number of open-ended and multi-category questions in our questionnaires. This is easier to say than do. Of course, another way to reduce interviewer effects is to increase the number of interviewers, thereby decreasing the average interviewer workload. This, though has scheduling and cost implications. Some comments on this alternative would have been useful.

While this analysis can help improve some aspects of interviewer training, it can not suggest in any reasonably definitive way what to do about the level and kind of interviewer training. The two data collection organizations of this paper seem to have two somewhat different approaches to training, with Westat providing significantly more training than the Census Bureau. I wish the sponsors and data collection organizations would do research on what works and what does not work in interviewer training. Whether more training provides better quality survey data? And what ought to be the relative balance between survey specific and general survey training?

While the authors articulate differences in models, population, and procedures, my feeling is that a substantial portion of the difference in magnitudes of the interviewer effects between the two surveys has nothing to do with the models, but with the combination of training offered and average workload size. An

organization effect may also be significant, but before anyone goes too far with that idea, I would review several papers on this subject(Cohen and Potter, 1990; Cohen, 1986; Cohen, 1982).

Finally, the results from the behavior coding study are not easily understood. The results are not highly correlated with the interviewer effects studies, but yet we all feel there ought to be some relationship. Perhaps, the small number of interviewers and the simple metric used were not capable of answering the questions posed by the authors.

Conclusion

While any review can raise a number of questions, and usually does, the important aspect of these papers is that they provide further evidence of the NSF commitment to study survey methods issues as a significant aspect of its data collection program. I congratulate the NSF staff for that and look forward to further presentations at future meetings of the American Statistical Association.

Bibliography

Brick, M.(1995). **Unit Response Rates in the 1995 National Household Education Survey**. Draft memorandum to the National Center for Education Statistics.

Bushery, J., Royce, D., and Kasprzyk, D.(1992). "The Schools and Staffing Survey: How Reinterview Measures Data Quality," **Proceedings of the Section on Survey Research Methods, American Statistical Association**, pp. 458-463, Alexandria, VA: American Statistical Association.

Citro, C. and Kalton, G. (1989). **Surveying the Nation's Scientists and Engineers : A Data System for the 1990's**, National Academy Press, Washington D.C.

Cohen, S.B.(1982). "Estimated Data Collection Organization Effect in the National Medical Care Expenditure Survey," **American Statistician**, 36, 337-341.

Cohen, S.B.(1986). "Data Collection Organization in the National Medical Care Utilization and Expenditure Survey," **Journal of Economic and Social Measurement**, 14, 367-378.

Cohen, S.B. and Potter, D.E.B.(1990). "Data Collection Organization Effects in the National Medical Expenditure Survey," **Journal of Official Statistics**, Vol. 6, No. 3, 275-294.

Jabine, T.B.(1994). **Quality Profile for SASS: Some Aspects of the Quality of Data in the Schools and Staffing Survey(SASS)**, NCES 94-340. Office of Educational Research and Improvement. Washington, D.C.: U.S. Government Printing Office.

Jabine, T.B.,King, K.E., and Petroni, R.J.(1990). **Survey of Income and Program Participation(SIPP): Quality Profile**, Washington, D.C.: U.S. Bureau of the Census.

Kasprzyk, D.(1992). "Discussion," **Proceedings of the Government Statistics Section, American Statistical Association**, pp. 100-102, Alexandria, VA: American Statistical Association.

Salvucci, S., Zhang, F., Monaco, D., Gruber, K., Scheuren, F.(1995). "Multivariate Modeling of Unit Nonresponse for SASS, 1990-91," **Proceedings of the Section on Survey Research Methods, American Statistical Association**, Alexandria, VA: American Statistical Association.